Computers Math. Applic. Vol. 24, No. 10, pp. 11–17, 1992
Printed in Great Britain.

0097-4943/92 $5.00 + 0.00

A NOTE ON PARAMETER ESTIMATION IN THE MULTIVARIATE BETA DISTRIBUTION

A. NARAYANAN
Decision and Information Systems, Indiana University
Bloomington, IN 47405, U.S.A.

(Received April 1991 and in revised form February 1992)

Abstract—This note describes a numerically feasible method for parameter estimation in the multivariate beta (Dirichlet) distribution through the method of maximum likelihood. The Dirichlet distribution is used in many applications, including brand choice modelling in marketing, investigating past vegetation changes in biology, etc. Examples involving the consumer brand choices of regular ground coffee and proportions of fossil pollen in trees and some simulation results are given.

1. INTRODUCTION

The multivariate beta distribution is an important distribution with wide ranging applications. One of these involves spurious correlations or correlations among proportions as investigated by Mosimann [1]. Biologists are generally interested in the correlations among proportions, since they are more easily observed then the actual numbers and represent the actual type of vegetation. In the analysis of fossil pollen described by Mosimann, counts of the frequency of occurrence of different kinds of pollen grains are made at various levels of core. Then, biologists are generally interested in making inferences about past vegetation changes from the data. If Y_i is a random variable giving value directly proportional to the number of pollen grains falling the area, then $X_i = Y_i / \sum_i Y_i$ gives the proportions of grains falling in the area. Suppose the Y_i's can be thought of as gamma variates independent of each other and the total, then the X_i's would follow a Dirichlet distribution. An example involving pollen grains from four kinds of trees is given in Section 4.

Another application concerns the modelling of consumer purchase behavior of nondurable items such as foods or toiletries [2]. A model for multibrand purchasing behavior is developed by Chatfield and Goodhardt [2], and theoretical justification is given or the "independent" case. Suppose there is a product field with k brands and let the random variable (Y_i) represent the average rate of purchase of brand i. Let $W = \sum_i Y_i$ represent the consumer's rate of buying of the product field as a whole. Then the joint distribution of $(X_1, X_2, \ldots, X_{k-1})$, where $(X_i = Y_i/W)$ represents the proportion of a consumer's total purchases devoted to brand i and follows a Dirichlet distribution. An example of brand choice using five brands of regular coffee is given in Section 4.

A third application would be modelling the activity times in a PERT (Program Evaluation and Review Technique) network. A PERT network has a collection of activities, and each activity is usually modelled as a random variable following a beta distribution. A Dirichlet distribution for the entire network immediately follows, since the marginal distributions of a Dirichlet is a beta. Using the properties of the Dirichlet distribution, we can see that each subnetwork also follows a Dirichlet distribution, and the distribution of the critical path will also follow a Dirichlet distribution. Monhor [3] uses the Dirichlet distribution for modelling the activity times of a PERT network and derives an upper bound for the completion time of the project. For a discussion of other applications, see [4].

In all the above applications, if the random vector $(X = X_1, X_2, \ldots, X_{k-1})$ follows a Dirichlet distribution, the joint density function is given by

$$f(X_1, \ldots, X_k) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k-1} X_j^{\alpha_j - 1} \left(1 - \sum_{j=1}^{k-1} X_j\right)^{\alpha_k - 1}, \qquad \mathbf{X} \in R^k,$$

where $R^k = \left[X_j \mid j = 1, \ldots, k;\ X_j > 0,\ \sum_{j=1}^{k} X_j = 1\right]$. This distribution is the multivariate extension of the 2-parameter beta distribution.

In this paper, we propose a numerically feasible method for the estimation of the parameters $(\alpha_1, \ldots, \alpha_k)$ of the multivariate beta distribution. In Section 2, we describe the Method of Moments (MM), and in Section 3, we describe the Method of Maximum Likelihood (ML). Finally, in Section 4, we show two examples and some simulation results.

2. METHOD OF MOMENTS

If $\mathbf{X} = \{X_1, \ldots, X_k\}$ follows a Dirichlet distribution with parameter vector $\alpha = \{\alpha_1, \ldots, \alpha_k\}$, then the first two population moments are [1]:

$$E(X_i) = \frac{\alpha_i}{\alpha}, \quad E(X_i^2) = \frac{\alpha_i (\alpha_i + 1)}{\alpha (\alpha + 1)}, \tag{2.1}$$

where $\alpha = \sum_{i=1}^{k} \alpha_i$. From an examination of the above equations, we see that there are k first order moments and k second order moments and thus, a total of $\binom{2k}{k}$ possible combinations of equations to solve for the k parameters. According to Fielitz and Myers [5], a symmetrical way of proceeding would be to choose the first $(k-1)$ first order equations and the first second order equation. The reason not to choose the k^{th} first order equation is that the k^{th} equation is a linear combination of the others and together they do not form an independent set of equations.

Define the sample moments as

$$X'_{1j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}; \qquad j = 1, \ldots, k,$$

$$X'_{21} = \frac{1}{n} \sum_{i=1}^{n} X_{i1}^2.$$

So

$$X'_{11} = \frac{1}{n} \sum_{i=1}^{n} X_{i1}. \tag{2.2}$$

Equating the suitable k population moments to sample moments as mentioned before and solving the set of equations, we get

$$\begin{aligned} \hat{\alpha}_i &= \frac{(X'_{11} - X'_{12}) X_{1i}}{X'_{21} - (X'_{11})^2}; \qquad i = 1, 2, \ldots, k-1, \\ \hat{\alpha}_k &= \frac{(X'_{11} - X'_{21}) \left(1 - \sum_{i=1}^{k-1} X'_{1i}\right)}{X'_{21} - (X'_{11})^2}. \end{aligned} \tag{2.3}$$

Derivations of these equations can be followed in [5]. These estimates can be used as starting values in the iterative solution of maximum likelihood methods.

3. METHOD OF MAXIMUM LIKELIHOOD

The idea of maximum likelihood estimates is to find those values of the parameters which most likely fit the observed data, i.e., maximizes the likelihood function (or log-likelihood function). Given a random sample of n observations from a Dirichlet distribution, the log likelihood function can be written as:

$$\log L = n\left\{\log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{j=1}^{k}\log\Gamma(\alpha_j)\right\} + n\sum_{j=1}^{k}(\alpha_j - 1)\log G_j, \tag{3.1}$$

where

$$G_j = \left[\prod_{i=1}^{n} X_{ij}\right]^{1/n}; \qquad j = 1, 2, \ldots, k-1, \quad \text{and} \quad G_k\left[\prod_{i=1}^{n}\left(\sum_{j=1}^{k-1} X_{ij}\right)\right]^{1/n}$$

are the geometric means of the k variables. Taking the derivatives of the log-likelihood function, the likelihood equations can be written as:

$$\frac{\partial \log L}{\partial \alpha_j} = n\,\Psi\left(\sum_{m=1}^{k}\alpha_m\right) - n\,\Psi(\alpha_j) + n\log G_j; \qquad j = 1, \ldots, k,$$

where $\Psi(\cdot)$ is the digamma function. The second partial and mixed partial derivatives are:

$$\begin{aligned}
\frac{\partial^2 \log L}{\partial \alpha_j^2} &= n\,\Psi'\left(\sum_{m=1}^{k}\alpha_m\right) - n\,\Psi'(\alpha_j); \qquad j = 1, \ldots, k, \\
\frac{\partial^2 \log L}{\partial \alpha_i\,\partial \alpha_j} &= n\,\Psi'\left(\sum_{m=1}^{k}\alpha_m\right).
\end{aligned} \tag{3.2}$$

The information matrix (**I**) is

$$\begin{aligned}
\mathbf{I} &= \{I_{ij}\} = -E\left[\frac{\partial^2 \log L}{\partial \alpha_i\,\partial \alpha_j}\right], \\
I_{ij} &= -n\,\Psi'\left(\sum_{m=1}^{k}\alpha_m\right), \qquad i \neq j, \\
I_{ii} &= n\,\Psi'(\alpha_i) - n\,\Psi'\left(\sum_{m=1}^{k}\alpha_m\right).
\end{aligned} \tag{3.3}$$

This matrix can be written as:

$$I = \mathbf{D} + \alpha\mathbf{a}\mathbf{b}', \tag{3.4}$$

where

$$\begin{aligned}
\mathbf{D} &= \operatorname{diag}\left[n\,\Psi'(\alpha_1), \ldots, n\,\Psi'(\alpha_k)\right], \\
\alpha &= -n\,\Psi'\left(\sum_{m=1}^{k}\alpha_m\right), \\
\mathbf{a} &= \mathbf{b} = 1'.
\end{aligned}$$

The variance-covariance matrix is obtained as the inverse of the Fisher information matrix (**I**) by using a well-known theorem [6, Theorem 8.3.3]. The variance-covariance matrix $\mathbf{V} = \{v_{ij}\}$ is thus found as:

$$\mathbf{V} = \mathbf{I}^{-1} = \mathbf{D}^* + \beta\,\mathbf{a}^*\,\mathbf{a}^{*\prime}, \tag{3.5}$$

where

$$\mathbf{D}^* = \operatorname{diag}\left[\frac{1}{n\,\Psi'(\alpha_1)}, \ldots, \frac{1}{n\,\Psi'(\alpha_k)}\right],$$

$$\mathbf{a}^* = \left[\frac{1}{\Psi'(\alpha_1)}, \ldots, \frac{1}{\Psi'(\alpha_k)}\right],$$

$$\beta = n\,\Psi'\left(\sum_{j=1}^{k}\alpha_j\right)\left[1 - \Psi'\left(\sum_{j=1}^{k}\alpha_j\right)\sum_{j=1}^{k}\frac{1}{\Psi'(\alpha_j)}\right]^{-1},$$

and $\Psi'(\bullet)$ is the trigamma function.

To numerically maximize the likelihood function (3.1), we will utilize a Newton-Raphson procedure which requires some estimate as an initial point of departure. For this purpose, the MM estimates of Section 2 or a suggestion given by Ronning [7] can be used. His suggestion is to set all $\alpha_j = \min\{X_{ij}\}$, $i = 1, \ldots, n$. These initial estimates prevent the α_j's from becoming a negative during the first few iterations. Given a set of initial estimates, Fisher's scoring method [8] can be used in the iterative scheme:

$$\begin{bmatrix}\hat{\alpha}_1\\ \vdots\\ \hat{\alpha}_k\end{bmatrix}_{(i)} = \begin{bmatrix}\hat{\alpha}_1\\ \vdots\\ \hat{\alpha}_k\end{bmatrix}_{(i-1)} + \begin{bmatrix}\operatorname{Var}(\hat{\alpha}_1) & \ldots & \operatorname{Cov}(\hat{\alpha}_1,\hat{\alpha}_k)\\ \vdots & \ddots & \vdots\\ \operatorname{Cov}(\hat{\alpha}_k,\hat{\alpha}_1) & \ldots & \operatorname{Var}(\hat{\alpha}_k)\end{bmatrix}_{(i-1)}\begin{bmatrix}g_1(\hat{\boldsymbol{\alpha}})\\ \vdots\\ g_k(\hat{\boldsymbol{\alpha}})\end{bmatrix}_{(i-1)},$$

where $\hat{\boldsymbol{\alpha}}_{[0]} = [\hat{\alpha}_{1(0)}, \ldots, \hat{\alpha}_{k(0)}]'$ are the initial estimates as described above. The usage of equation (3.5) makes the computation of the variance-covariance matrix easier and avoids inverting the information matrix at every iteration of the Newton-Raphson algorithm. The availability of numerical routines to compute the digamma function [9] and trigamma function [10] for the elements of the variance-covariance matrix makes the ML method computationally feasible.

The test for convergence of the method can be done using a quadratic form of the gradient vector. Consider the statistic

$$S = [g_1(\hat{\boldsymbol{\alpha}}), \ldots, g_k(\hat{\boldsymbol{\alpha}})]\begin{bmatrix}\operatorname{Var}(\hat{\alpha}_1) & \ldots & \operatorname{Cov}(\hat{\alpha}_1,\hat{\alpha}_k)\\ \vdots & \ddots & \vdots\\ \operatorname{Cov}(\hat{\alpha}_k,\hat{\alpha}_1) & \ldots & \operatorname{Var}(\hat{\alpha}_k)\end{bmatrix}\begin{bmatrix}g_1(\hat{\boldsymbol{\alpha}})\\ \vdots\\ g_k(\hat{\boldsymbol{\alpha}})\end{bmatrix}.$$

This test statistic can be shown to be distributed as a chi-square random variable with k degrees of freedom [11]. Although this is a large sample test, it proves to be quite valuable even in samples of moderate sizes. The iteration is continued until S becomes less than $\chi^2_k(\nu)$ for a fixed ν in the lower tail of the chi-square distribution with k degrees of freedom. Such an approach was also used by Choi and Wette [12] in a similar situation for the gamma distribution.

4. EXAMPLE AND RESULTS

We initially show an example of brand choice for regular ground coffee from store and panel records taken from four Kansas City supermarkets. The data were collected by Selling Areas-Marketing, Inc. (SAMI) and were kindly made available for this research.

The coffee market is generally divided into six product groups, depending upon whether they are ground/instant, caffeinated/decaffeinated and, within instant, freeze dried/non-freeze dried. To obtain a relatively homogeneous market, we chose regular ground coffee.

Another issue is package size. In Kansas City, the popular sizes are 1 pound and 3 pounds. We shall call these "small" and "large" (included in "small" are 13 and 14 ounce sizes, which are advertised as producing the same number of cups as 1 pound). Different brand sizes are perceived differently, both by consumers and producers, and we model brand sizes. We chose three "small" sizes (brands A, B and C) and two "large" sizes (brands D and E). Purchase rates for each of the five brands were computed across a group of consumers, and relative frequencies (probabilities of purchase) for each brand were estimated. After deleting some observations for incomplete data, 28 vectors of probabilities were obtained for all five brands. When a Dirichlet distribution

Table 1. Maximum likelihood estimates and variance-covariance matrix of the estimated parameters for the coffee data.

Maximum likelihood estimates of the parameters

$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$
6.446	3.775	2.284	2.329	1.239

Variance-covariance matrix of the estimated parameters

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$
$\hat{\alpha}_1$	0.883				
$\hat{\alpha}_2$	0.374	0.322			
$\hat{\alpha}_3$	0.207	0.115	0.127		
$\hat{\alpha}_4$	0.210	0.116	0.645	0.130	
$\hat{\alpha}_5$	0.091	0.050	0.028	0.028	0.040

was fitted to the data, the following maximum likelihood estimates and the variance-covariance matrix were obtained (Table 1).

Using these estimates, a $100(1-\alpha)\%$ confidence interval of the form $\hat{\alpha}_j \pm Z_{\alpha/2} V(\hat{\alpha}_j)$ can be constructed. For example, a 95% confidence interval on α_1 is (4.72, 8.18).

The next example is from biology and the data are taken from [1]. Frequency of occurrence of four (pine, fir, oak and alder) types of grains falling under different levels of core are given. An attempt is then made by biologists to reconstruct past vegetation changes in the area from which the core was taken. If each frequency count is represented by Y_i, then the proportions $Y_i / \sum_i Y_i$ can be considered to follow a multivariate beta distribution. The parameter estimates and the associated covariance matrix needed for such inferences are given in Table 2.

Table 2. Maximum likelihood estimates and variance-covariance matrix of the parameters for fossil data.

Maximum likelihood estimates of the parameters

$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
36.734	1.232	3.9780	1.794

Variance-covariance matrix of the parameters

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
$\hat{\alpha}_1$	24.252			
$\hat{\alpha}_2$	0.527	0.031		
$\hat{\alpha}_3$	2.260	0.051	0.302	
$\hat{\alpha}_4$	0.872	0.020	0.084	0.065

Confidence intervals for the parameter or linear combinations of parameters can be obtained in a straightforward manner.

A simulation study was also conducted to analyze the small sample properties of the algorithm. The results of the simulation study for selected values of the parameters and different sample sizes are shown in Table 3.

Parameter values were selected such that they represented a combination of small and large values and also a mixture of equal values versus a mixture of values of different sizes. Values in the table are average estimates ($\hat{\alpha}_i$) and mean square errors (MSE(α_i)) over 100 replications. As can be seen from the table, for each combination of parameter values, the mean square error (MSE)

Table 3. Simulation results for Dirichlet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and associates mean squared errors over 100 replications.

α_1	α_2	α_3	α_4	N	α_1	α_2	α_3	α_4	MSE($\hat{\alpha}_1$)	MSE($\hat{\alpha}_2$)	MSE($\hat{\alpha}_3$)	MSE($\hat{\alpha}_4$)
0.5	0.5	1.0	1.0	10	0.26	0.63	1.66	1.54	0.93	0.78	2.43	1.76
				25	0.52	0.46	1.05	1.10	0.03	0.08	0.15	0.15
				40	0.50	0.52	1.06	1.05	0.02	0.02	0.12	0.10
3.0	6.0	9.0	12.0	10	2.95	7.96	12.49	17.0	6.02	33.33	82.91	143.34
				25	2.99	6.26	9.46	12.58	0.53	2.51	5.76	11.79
				40	3.10	6.24	9.24	12.40	0.31	11.384	2.71	5.12
5.0	5.0	5.0	5.0	10	3.55	6.95	7.09	7.19	108.72	45.57	38.07	37.57
				25	5.27	5.32	5.28	5.29	1.75	1.63	1.76	1.70
				40	5.17	5.24	5.21	5.23	1.08	1.07	1.08	1.00
12.5	10.0	5.0	2.5	10	12.94	13.40	6.72	3.52	108.72	45.57	38.07	37.57
				10	12.87	10.80	5.41	2.66	35.04	12.68	3.64	1.70
				40	13.00	10.69	5.39	2.66	9.95	5.88	1.84	0.48
0.5	2.5	13.0	20.0	10	0.88	3.77	28.72	39.28	2.05	38.30	959.71	1929.85
				25	0.43	2.82	19.86	26.15	0.11	4.80	149.38	240.44
				40	0.50	16.41	21.99	0.03	0.03	1.03	30.68	51.70

decreases as the sample size increases. Also within each combination of parameter values, the MSE is higher for larger parameter values. A third and important finding is that the structure of parameters (whether they are equal or proportional) does not matter, at least in terms of mean square errors. For example, in the third experiment, $(\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 5.0)$, the MSE for α_3 is about the same as the MSE for α_3 in the fourth experiment, $(\alpha_1 = 12.5,\ \alpha_2 = 10.0,\ \alpha_3 = 5.0,\ \alpha_4 = 2.5)$. Other combinations (not reported here) gave similar results.

The mean square errors reported in the table are a combination of variance and bias. In small sample sizes $(n \leq 10)$, the bias component is considerably large and decreases and the sample size increases. This kind of behavior is characteristic of maximum likelihood estimates.

It should be noted that the results given here are from a small Monte Carlo study, and a larger study is needed to come to general conclusions. However, these results show that unless the sample sizes are too small $(n \leq 10)$, maximum likelihood estimates do fairly well and carry over their large sample properties even to samples of moderate size.

To conclude, we have proposed a method for maximum likelihood estimation of the parameters of the multivariate beta distribution, which is used in a number of applications. This method is also useful in estimating the covariance matrix of the parameters. Simulation results indicate that the method has good large and small sample properties. (A FORTRAN program for the algorithm can be found in [13].)

REFERENCES

1. J.E. Mosimann, On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions, *Biometrika* **49** (1/2), 65–82 (1962).
2. C. Chatfield and G. Goodhardt, Results concerning brand choice, *Journal of Marketing Research* **12** (February), 110–113 (1975).
3. D. Monhor, An approach to PERT: Application of Dirichlet distribution, *Optimization* **18**, 113–118 (1987).
4. N.L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley, New York, (1972).
5. B. Fielitz and B.L. Myers, Estimation of parameters in the beta distribution, *Decision Sciences* **6** (1), 1–13 (1975).
6. F.A. Graybill, *Matrices with Applications in Statistics*, 2nd edition, Wadsworth, California, (1983).
7. G. Ronning, Maximum likelihood estimation in Dirichlet distribution, *Journal of Statistical Computation and Simulation* **32**, 215–221 (1989).
8. C.R. Rao, *Advanced Statistical Methods in Biometric Research*, John Wiley, New York, (1952).
9. J.M. Bernardo, Algorithm AS 103: PSI (Digamma) function, *Applied Statistics* **25**, 315–317 (1976).
10. B.E. Schneider, Algorithm AS 121: Trigamma function, *Applied Statistics* **27**, 97–98 (1978).
11. R. Serfling, *Approximation Theorems in Mathematical Statistics*, John Wiley, New York, (1980).

12. S.C. Choi and R. Wette, Maximum likelihood estimation of the parameters of the gamma distribution and their bias, *Technometrics* **11**, 683–689 (1967).
13. A. Narayanan, Algorithm AS 266: Maximum likelihood estimation of the parameters of the Dirichlet distribution, *Applied Statistics* **40** (2), 365–374 (1991).